Revealing the Traces of JPEG Compression Anti-Forensics

Giuseppe Valenzise, *Member, IEEE*, Marco Tagliasacchi, *Member, IEEE*, and Stefano Tubaro, *Member, IEEE*

***Abstract*—Due to the lossy nature of transform coding, JPEG introduces characteristic traces in the compressed images. A forensic analyst might reveal these traces by analyzing the histogram of discrete cosine transform (DCT) coefficients and exploit them to identify local tampering, copy-move forgery, etc. At the same time, it has been recently shown that a knowledgeable adversary can possibly conceal the traces of JPEG compression, by adding a dithering noise signal in the DCT domain, in order to restore the histogram of the original image. In this paper, we study the processing chain that arises in the case of JPEG compression anti-forensics. We take the perspective of the forensic analyst, and we show how it is possible to counter the aforementioned anti-forensic method revealing the traces of JPEG compression, regardless of the quantization matrix being used. Tests on a large image dataset demonstrated that the proposed detector was able to achieve an average accuracy equal to 93%, rising above 99% when excluding the case of nearly lossless JPEG compression.**

***Index Terms*—Anti-forensics, digital image forensics, JPEG compression.**

I. INTRODUCTION

THE availability of low-cost digital cameras, together with the widespread adoption of multimedia sharing platforms, have made the acquisition and the dissemination of digital images a virtually costless job. For this reason, images have become a popular and easy means to convey information. At the same time, producing photorealistic forgeries of original content has become a rather simple task, even for non professional users. In many cases, these forgeries might be innocuous. However, in some circumstances they could be used for malicious purposes, e.g., when the doctored images are employed as evidence in courtrooms, or as material for propaganda, etc. In those cases, forgeries may aim at discrediting somebody's reputation or at altering facts to influence the public opinion. In order to limit the hazards connected to misuse of digital images, in the past few years a variety of digital image forensic techniques have been proposed by the forensic community [3], [4]. Differently from *watermarking* or *hashing*, these methods do not rely on extrinsic information embedded into the image at the moment of acquisition, or received by a secure server upon demand. In fact, the questioned image is typically the only source of information available to the forensic analyst. Therefore, forensic techniques analyze the image content in order to find traces left by specific acquisition, coding or editing operations, which could be telltale of malicious tampering. These traces include, e.g., footprints left by the camera sensor noise [5]; coding [6], [7]; resampling [8]; cropping [9]; and point-wise processing [10].

The footprints left by JPEG compression play an important role in detecting possible forgeries, since JPEG is by far the most widely used image compression standard. To achieve lossy compression, a JPEG encoder quantizes each discrete cosine transform (DCT) coefficient of an image to multiples of a quantization step size, specified by the JPEG quantization matrix. When an image is decoded, the distribution of reconstructed DCT coefficients differs from the original, i.e., it exhibits a characteristic comb-like shape, which might reveal the original quantization matrix [6]. This fact enables several forensic analysis tasks, including the identification of which camera took a picture [11], or the detection of double JPEG compression [12]–[14]. Furthermore, localized evidence of double compression can reveal "copy-move" forgeries, in which an adversary copies portions of other images into the doctored picture, before resaving the result as JPEG [9], [15]–[17].

Given the relevance of JPEG compression footprints in image forensics, a natural question arises about whether these traces could be concealed by a knowledgeable adversary. Indeed, Stamm *et al.* [18] have shown that the statistical footprints of JPEG compression can be removed by adding a properly designed dithering noise signal to the quantized DCT coefficients of a JPEG-compressed image. The distribution of the dithering noise signal is such that the resulting coefficients are approximately distributed as those of the uncompressed original image. Using this technique, the authors of [18] have also demonstrated that many of the aforementioned forensic techniques based on JPEG footprints can be fooled [19]. Nevertheless, this anti-forensic approach is not exempt from leaving behind traces of its own.

In this work, we build on the observation that the anti-forensic dither is a noisy signal which cannot replace the image content lost during quantization. As that, it introduces visible distortion in the attacked image, which appears as a characteristic grainy

Manuscript received July 31, 2012; revised October 24, 2012; accepted December 07, 2012. Date of publication December 20, 2012; date of current version January 09, 2013. The material in this paper was presented in part at the International Conference on Acoustics, Speech, and Signal Processing, Prague, Czech Republic, May 2011, and the International Conference on Image Processing, Brussels, Belgium, September 2011. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. C.-C. Jay Kuo.

G. Valenzise is with the Laboratoire Traitement et Communication de l'Information (LTCI-UMR 5141) of the CNRS at Télécom ParisTech, 75014 Paris, France (e-mail: giuseppe.valenzise@telecom-paristech.fr).

M. Tagliasacchi and S. Tubaro are with the Dipartimento di Elettronica e Informazione, Politecnico di Milano, 32 20133-Milano, Italy (e-mail: marco.tagliasacchi@polimi.it; stefano.tubaro@polimi.it).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIFS.2012.2234117

noise that allows to discriminate attacked images from original uncompressed images. In our previous work [1] we have analyzed these traces in terms of distortion introduced in the tampered image. Based on that, we take here the perspective of the forensic analyst and show how he can effectively counter the anti-forensic measures adopted by an adversary operating according to [18]. To this end, we extend our previous work [2] to the general and more challenging scenario in which the quantization matrix template is unknown to the forensic analyst. The core of the proposed detector consists in recompressing the questioned image by varying the coding conditions and analyzing the amount of grainy noise left by the adversary. We design an anti-forensics detector which only needs to change, at each compression round, a pair of properly selected DCT coefficients. This recompress-and-observe paradigm is inspired to similar methods presented in the literature, which exploit the idempotency property of quantization in order to, e.g., estimate the quality factor in JPEG compressed images [20], exposing forgeries [21] and to identify the video codec [22]. However, to the best of the authors' knowledge, it has never been applied to the problem of detecting JPEG compression anti-forensics before. Our experiments demonstrate that it is possible to correctly detect attacked images with an accuracy equal to 93%, which rises above 99% when excluding the case of nearly lossless JPEG compression. These results indicate that removing JPEG compression footprints is not as simple as previously thought, since the process of footprint removal inevitably introduces new traces in the doctored image.

We observe that there is a tight relationship between steganalysis and forensic analysis. Indeed, some methods originally developed to detect stego images can be used for the problem at hand, by considering the uncompressed original image as the cover image, and the decompressed JPEG image with added the anti-forensic dithering signal as the stego image. This is, e.g., the approach recently taken in [23], in which JPEG compression anti-forensics is countered by means of calibration features, originally proposed in the field of steganalysis [24]. In addition to [23], we also include in our experiments the detector based on SPAM (Subtractive Pixel Adjacency Matrix) features [25], which were proposed to detect steganographic methods that embed in the spatial domain by adding a low-amplitude independent stego signal. Notice that the above mentioned steganographic approaches are effective to detect whether an additive noise signal has been added to the cover image (thus including the anti-forensic dither as a special case). Conversely, the approach proposed in this paper targets the specific case of JPEG compression anti-forensic dither, based on the idempotency property of quantization. This entails a lower false positive rate when other kinds of noise are introduced into the image without malicious purposes.

Note that we assume that the adversary saves the attacked image according to a lossless compression format, consistently with the previous literature [18]. This is a reasonable standpoint, since the ultimate goal of the adversary is to conceal the traces of lossy compression. In some cases, though, the adversary might want to save the attacked image with JPEG or other lossy compression formats, presumably at quality higher than the one of the original JPEG compression. Revealing the traces of JPEG compression anti-forensics in the case of double [13][26] (or multiple [27]) compression is intuitively harder, and it is left to future investigation.

The rest of the paper is organized as follows. Section II reviews the basics of JPEG compression and summarizes the anti-forensic technique described in [18]. For convenience, in Section III we illustrate the study appeared in our previous work [1], where we analyzed the mean-square-error distortion introduced by anti-forensic dithering. This is instrumental in setting the theoretical basis for the detection of anti-forensic dithering based on recompression, which is described in Section IV and experimentally validated in Section V, for the two cases of known and unknown quantization matrix template. Finally, Section VI concludes the paper.

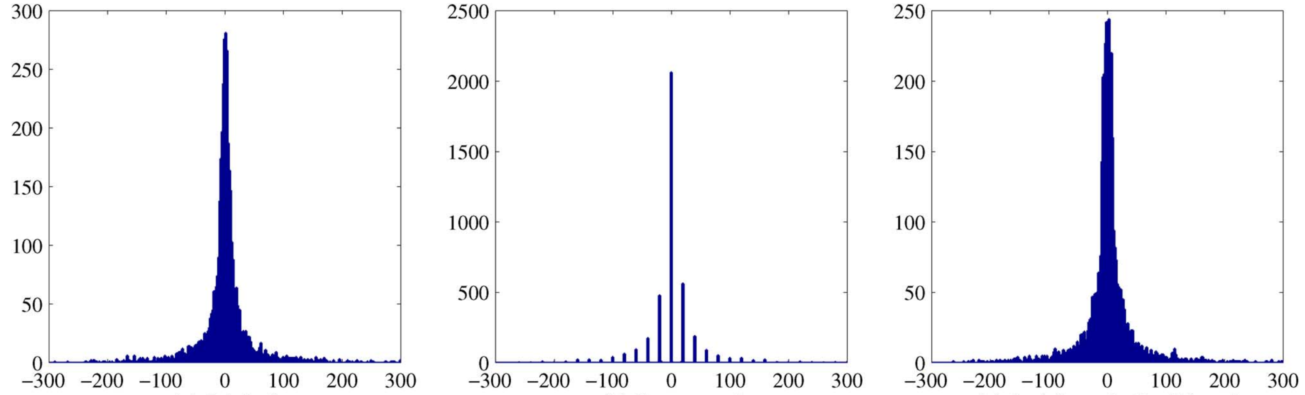
II. Background

In this section we start reviewing the basics of JPEG compression and the corresponding footprints. Then, we summarize the anti-forensic technique described in [18]. Without loss of generality, we consider JPEG compression for gray-scale images. However, the very same principles apply to the luminance and chrominance channels of color images.

A. JPEG Compression and JPEG Compression Footprints

In the JPEG compression standard, the input image is first divided into B nonoverlapping pixel blocks of size 8×8. For each block, the two-dimensional discrete cosine transform (DCT) is computed. Let X_i^b, $1 \leq b \leq B$, $1 \leq i \leq 64$, denote the i-th transform coefficient of the b-th block according to some scanning order (e.g., zig-zag). That is, there is a one-to-one mapping $i \leftrightarrow (r, s)$ between the index i and the position (r, s), $1 \leq r, s \leq 8$, of a coefficient within a DCT block. Let $\mathbf{X}_i = [X_i^1 \ldots, X_i^b]^T$ denote the set of DCT coefficients of the i-th subband. Each DCT coefficient X_i^b, $1 \leq i \leq 64$, is quantized with a quantization step size q_i. The set of q_i's forms the quantization matrix $\mathbf{Q}$, which is *not* specified by the standard. In many JPEG implementations, it is customary to define $\mathbf{Q}$ as a scaled version of a template matrix, by adjusting a (scalar) quality factor Q. This is the case, for instance, of the quantization matrices adopted by the Independent JPEG Group (IJG) [28], which are obtained by properly scaling the image-independent quantization matrices suggested in Annex K of the JPEG standard [29]. The quantization levels W_i^b are obtained from the original coefficients X_i^b as $W_i^b = \text{round}(X_i^b / q_i)$. The quantization levels are entropy coded and written in the JPEG bitstream. When the bitstream is decoded, the DCT values are reconstructed from the quantization levels as $\tilde{X}_i^b = q_i W_i^b$. Then, the inverse DCT is applied to each block, and the result is rounded and truncated in order to take integer values on [0,255].

Due to the quantization process, the dequantized coefficients $\tilde{X}_i^b$ can only assume values that are integer multiples of the quantization step size q_i. Therefore, the histogram of dequantized coefficients of the i-th DCT subband, i.e., $\tilde{\mathbf{X}}_i = [\tilde{X}_i^1, \ldots, \tilde{X}_i^B]$, is comb-shaped with peaks spaced apart by q_i. This is depicted in Fig. 1, which shows the histogram of transform coefficients in the (2,1) DCT subband before (Fig. 1(a)) and after (Fig. 1(b)) JPEG compression. The process



(a) Original (b) Compressed (c) Anti-forensically dithered

Fig. 1. (a) Histogram of transform coefficients in the (2,1) DCT subband for the *Lenna* image. (b) Due to quantization, the coefficients in the JPEG-compressed image have a comb-shaped distribution. (c) The anti-forensic technique in [18] enables us to approximately restore the original distribution, thus removing JPEG-compression footprints.

of rounding and truncating the decompressed pixel values perturbs the comb-shaped distribution of $\tilde{\mathbf{X}}_i$. However, the DCT coefficient values typically remain tightly clustered around integer multiples of q_i. Hereafter, we refer to this characteristic comb shape of the DCT coefficients' histogram as *JPEG compression footprint*, as it reveals that: a) a quantization process has occurred; and b) which was the original quantization step size [6].

B. JPEG Compression Anti-Forensics

The work in [18] proposes to conceal the traces of JPEG compression by filling the gaps in the comb-shaped distribution of $\tilde{\mathbf{X}}_i$ by adding a dithering, noise-like, signal $\mathbf{N}_i$ in such a way that the distribution of the dithered coefficients $\mathbf{Y}_i = \tilde{\mathbf{X}}_i + \mathbf{N}_i$ approximates the original distribution of $\mathbf{X}_i$. The original AC coefficients $(2 \le i \le 64)$ are typically assumed to be distributed according to the Laplacian distribution [30]:

$$f_{\mathbf{X}_i}(x) = \frac{\lambda_i}{2} e^{-\lambda_i |x|}, \tag{1}$$

where the decay parameter λ_i typically takes values between 10^{-3} and 1 for natural imagery. In practice, only the JPEG-compressed version of the image is available, and the original AC coefficients $\mathbf{X}_i$ are unknown. Therefore, the parameter λ_i in (1) must be computed from the quantized coefficients $\tilde{\mathbf{X}}_i$, e.g., using the maximum-likelihood method in [31], which will result in an estimated parameter $\hat{\lambda}_i$.

According to [18], in order to remove the statistical traces of quantization in $\tilde{\mathbf{X}}_i$, the dithering signal $\mathbf{N}_i$ needs to be designed in such a way that its distribution depends on whether the corresponding quantized coefficients $\tilde{\mathbf{X}}_i$ are equal to zero. That is, for DCT coefficients quantized to zero:

$$f_{\mathbf{N}_i}(n|\tilde{X}_i = 0) = \begin{cases} \frac{1}{c_0} e^{-\hat{\lambda}_i |n|} & \text{if } -\frac{q_i}{2} \le n < \frac{q_i}{2} \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where $c_0 = (2/\hat{\lambda}_i)(1 - e^{-\hat{\lambda}_i q_i/2})$. Conversely, for the other coefficients

$$f_{\mathbf{N}_i}(n|\tilde{X}_i = x) = \begin{cases} \frac{1}{c_1} e^{-\mathrm{sgn}(x)\hat{\lambda}_i (n + q_i/2)} & \text{if } -\frac{q_i}{2} \le n < \frac{q_i}{2} \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where $c_1 = (1/\hat{\lambda}_i)(1 - e^{-\hat{\lambda}_i q_i})$. Note that the value of the DCT coefficient $\tilde{X}_i$ enters the definition of the p.d.f. in (3) only through its sign. For some DCT subbands, all the coefficients may be quantized to zero, and $\hat{\lambda}_i$ cannot be determined. In those cases, the authors of [18] suggest to leave the reconstructed coefficients unmodified, i.e., $\mathbf{Y}_i = \tilde{\mathbf{X}}_i$.

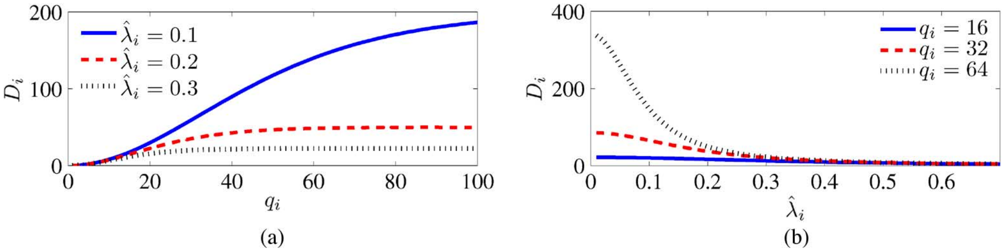
As for the DC coefficients, there is no general model for representing their distribution. Hence, the anti-forensic dithering signal for the DC coefficient $(i = 1)$ is sampled from the uniform distribution

$$f_{\mathbf{N}_1}(n) = \begin{cases} \frac{1}{q_i} & \text{if } -\frac{q_i}{2} \le n < \frac{q_i}{2} \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Fig. 4 illustrates that the anti-forensic technique enables to approximately restore the original Laplacian distribution, thus removing JPEG-compression footprints.

III. Analysis of the Distortion Introduced by Anti-Forensic Dithering

The addition of the anti-forensic dither illustrated in the previous section corresponds to injecting a noise-like signal in the pixel domain. As a result, the dithered image is distorted with respect to the JPEG-compressed image. In this section, we characterize analytically the distortion in the DCT domain, showing that it is a function of both the distribution of the original transform coefficients and the quantization step size. We arrive to the conclusion that the energy of anti-forensic dithering is concentrated in the middle DCT frequencies, thus resulting in a grainy noise in the spatial domain. We leverage this fact in Section IV-B to select a proper set of DCT coefficients to analyze in order to detect JPEG anti-forensics. Next, in Section III-B we analyze the effect of requantizing the dithered coefficients in a DCT subband using different quantization step sizes. We observe that requantizing the dithered coefficients with the original JPEG quantization step annihilates completely the anti-forensic noise. We build on this observation to detect the traces left by JPEG compression anti-forensics in Section IV.



(a) (b)

Fig. 2. MSE distortion D_i introduced by anti-forensic dithering for (a) different values of the decay parameter $\hat{\lambda}_i$; (b) different values of quantization step size q_i.

A. Characterization of Anti-Forensic Dithering Energy

The mean-square-error (MSE) distortion $\hat{D}_i$ between the JPEG-compressed coefficients $\tilde{\mathbf{X}}_i$ and the dithered coefficients $\mathbf{Y}_i$ in the i-th subband can be measured directly in the DCT domain, since the transform is orthonormal. That is,

$$\hat{D}_i = \frac{1}{B}\sum_{b=1}^{B}\left(Y_i^b - \tilde{X}_i^b\right)^2 = \frac{1}{B}\sum_{b=1}^{B}\left(N_i^b\right)^2. \qquad (5)$$

Since the distribution of the dithering signal $\mathbf{N}_i$ is known from (2)–(4), it is possible to obtain an analytical expression of the expected value $D_i = E[\hat{D}_i]$:

$$D_i = \sum_{k=-\infty}^{+\infty} \Pr(\tilde{X}_i = kq_i) \int_{-q_i/2}^{+q_i/2} x^2 f_{\mathbf{N}_i}(x|\tilde{X}_i = kq_i)dx, \quad (6)$$

where $\Pr(\tilde{X}_i = kq_i)$ represents the probability mass function of the quantized DCT coefficients. For AC coefficients, (6) can be rewritten according to the definitions given in (2), (3). That is,

$$D_i = m_i^0 D_i^0 + \left(1 - m_i^0\right) D_i^1, \quad \text{for} \quad 1 < i \leq 64 \qquad (7)$$

where

$$D_i^0 = \int_{-q_i/2}^{+q_i/2} x^2 f_{\mathbf{N}_i}(x|\tilde{X}_i = 0)dx, \qquad (8)$$

$$D_i^1 = \int_{-q_i/2}^{+q_i/2} x^2 f_{\mathbf{N}_i}(x|\tilde{X}_i = kq_i)dx, \qquad (9)$$

and $m_i^0 = 1 - e^{-\hat{\lambda} q_i/2}$ is the fraction of coefficients quantized to zero.

For DC coefficients, the mean square error D_1 is equal to that of a uniform scalar quantizer, i.e., $D_1 = q_1^2/12$. Instead, for AC coefficients, an expression can be found in closed form by solving the integrals in (8) and (9), as a function of the quantization step size and the parameter of the Laplacian distribution. That is:

$$D_i^0(q_i, \hat{\lambda}_i) = \frac{\hat{\lambda}_i^2 q_i^2 + 4\hat{\lambda}_i q_i + 8\left(1 - e^{\frac{\hat{\lambda}_i q_i}{2}}\right)}{4\hat{\lambda}_i^2\left(1 - e^{\frac{\hat{\lambda}_i q_i}{2}}\right)}, \qquad (10)$$

$$D_i^1(q_i, \hat{\lambda}_i) = \frac{1}{4}q_i^2 + 2\frac{q_i}{\hat{\lambda}_i}\left(\frac{1}{1 - e^{\hat{\lambda}_i q_i}} - \frac{1}{2}\right) + \frac{2}{\hat{\lambda}_i^2}. \quad (11)$$

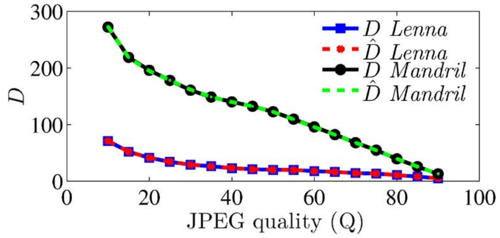


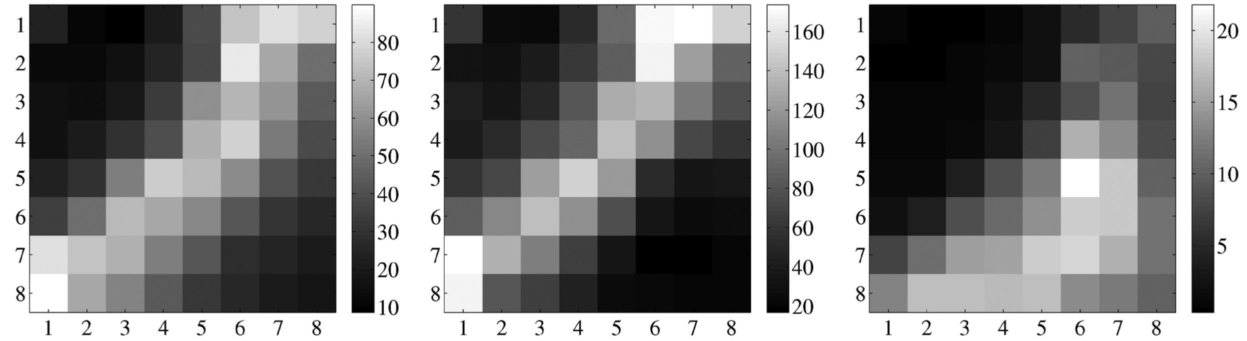
Fig. 3. MSE distortion D for two images characterized by different smoothness. The distortion introduced when removing JPEG footprints from images characterized by high-frequency textures is in general higher than for smooth images.

Fig. 2(a) shows the MSE distortion D_i as a function of q_i, for different values of $\hat{\lambda}_i$. As a general consideration, the distortion introduced by the anti-forensic dither gets larger as the quantization step size increases. Indeed, a larger value of q_i implies a wider spacing between the peaks in the comb-shaped distribution of $\tilde{\mathbf{X}}_i$. Thus, a larger amount of noise needs to be added to restore the original coefficient distribution.

The growth of the mean square error D_i depends also on the value of $\hat{\lambda}_i$, as illustrated in Fig. 2(b). A larger $\hat{\lambda}_i$ in the Laplacian model (1) results in DCT coefficients which are more clustered around zero (i.e., with smaller energy). When $\hat{\lambda}_i$ is sufficiently large, all coefficients fall into the zero bin of the quantizer (i.e., $m_i^0 = 1$ in (7)). Therefore, no anti-forensic noise is added, and the distortion is exactly zero.

As an example, Fig. 3 shows the MSE distortion $\hat{D} = (1/64)\sum_{i=1}^{64} \hat{D}_i$ as a function of JPEG compression quality Q, when the IJG quantization matrices are used. We consider two images with different content characteristics, *Lenna* and *Mandril*. Visual inspection reveals that *Lenna* is smoother than *Mandril*. In the DCT domain, larger values of $\hat{\lambda}_i$ are observed for *Lenna*, especially at high frequency. Therefore, the MSE distortion introduced in *Lenna* is smaller than in *Mandril*. Fig. 3 also demonstrates that the analytical model describing the expected distortion D introduced by dithering provides an accurate match of the measured MSE $\hat{D}$.

We characterize the distribution of the MSE distortion due to the anti-forensic dither in the DCT domain in order to understand which DCT frequencies are affected more by dithering. This is useful in order to tune the parameters for the anti-forensic detector explained in Section IV-B. In order to exploit frequency masking and increase coding efficiency, the quantization step sizes q_i's are generally larger at higher frequencies. Hence, we expect larger values of $\hat{D}_i$ in those DCT subbands corresponding to higher frequencies. On the



(a) Average (b) $Q = 30$ (c) $Q = 90$

Fig. 4. MSE distortion $\hat{D}_i$ in the 64 DCT subbands for different JPEG quality factors Q, averaged over the collection of images in the UCID dataset [32]. (a) Average MSE distortion over several quality factors in the range [30, 95]. (b) MSE distortion at $Q = 30$. (c) MSE distortion at $Q = 90$.

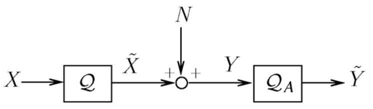


Fig. 5. Scheme of the requantization of a DCT coefficient.

other hand, high-frequency components have lower energy (higher $\hat{\lambda}_i$) due to the piecewise smoothness of natural images. As a result, we typically observe larger values of the MSE distortion at intermediate frequencies. This is illustrated in Fig. 4(a), which shows $\hat{D}_i$ averaged over all the 8×8 blocks of the images in the UCID color image database [32], compressed using JPEG at different quality factors $Q \in [30, 95]$ with the IJG quantization matrices. We notice that $\hat{D}_i$ is not uniformly distributed across the DCT subbands, and it is concentrated at medium frequencies. Indeed, the distribution depends on both the image content and the quantization matrix employed. As a further example, Fig. 4(b) and Fig. 4(c) show the average MSE distortion when all images in the dataset are compressed at a quality factor equal to, respectively, $Q = 30$ and $Q = 90$. When the quality of JPEG compression increases: i) the overall amount of distortion decreases (see the different scale being used); ii) the distribution of the MSE distortion is shifted towards DCT subbands that correspond to higher frequencies.

B. Effect of Requantization on the Dithered Image

The distribution of the anti-forensic dither in the i-th DCT subband is designed in such a way that it is nonzero only in the interval $[-(q_i/2), q_i/2)$, where q_i is the corresponding quantization step size (cfr. (2)–(4)). Based on this observation, we show that the anti-forensic noise can be completely canceled if the dithered image is requantized using the same quantization matrix used in the original JPEG compression step.

For clarity of illustration, we start considering a *single* coefficient in one DCT subband, requantized according to the scheme illustrated in Fig. 5. In order to simplify the notation, we drop the subband index, e.g., $q_i = q$. Let X denote the value of a DCT coefficient in the original (uncompressed) image. During JPEG compression, X is quantized using a uniform quantizer $\mathcal{Q}$ with quantization step size q, thus producing $\tilde{X}$. In order to remove the traces of quantization, an adversary adds the dithering signal N according to [18], thus producing Y. From the discussion in Section II-B, the net result is that the p.d.f. $f_{\mathbf{Y}}(y)$ is indistinguishable from $f_{\mathbf{X}}(x)$. Then, the dithered coefficient Y is requantized with a uniform quantizer $\mathcal{Q}_A$ with quantization step size q_A, producing the new coefficient $\tilde{Y}$. We are interested in computing the MSE distortion, $D_A(q_A)$, between $\tilde{X}$ and $\tilde{Y}$. That is,

$$\begin{aligned} D_A(q_A) &= E\left[(\tilde{X}-\tilde{Y})^2\right] \\ &= \sum_{k=-\infty}^{+\infty} p_k \left[\int_{-\frac{q}{2}}^{+\frac{q}{2}} \left(\tilde{x}_k - \mathcal{Q}_A(\tilde{x}_k+n)\right)^2 f_{\mathbf{N}}(n)dn\right], \quad (12) \end{aligned}$$

where $\tilde{x}_k = kq$, the expectation is taken with respect to the joint distribution of $\tilde{X}$ and N, $f_{\mathbf{N}}(n)$ is the probability density function of the dithering noise as in (2)–(4), and

$$p_k = \int_{kq-\frac{q}{2}}^{kq+\frac{q}{2}} f_X(x)dx \quad (13)$$

is the probability of the original coefficient X to fall in the k-th quantization bin. Notice that the output of $\mathcal{Q}_A$ assumes values at integer multiples of q_A. Therefore, after a change of variables, (12) can be written as

$$D_A(q_A) = \sum_{k=-\infty}^{+\infty} p_k \left[\sum_{h=-\infty}^{+\infty} (kq - hq_A)^2 p_{h|k}\right], \quad (14)$$

where

$$p_{h|k} = \int_{hq_A-\frac{q_A}{2}}^{hq_A+\frac{q_A}{2}} f_N(n-kq)dn \quad (15)$$

is the probability of quantizing a dithered sample to the h-th bin of $\mathcal{Q}_A$, given that the original sample was quantized to the k-th bin of $\mathcal{Q}$. This is illustrated in Fig. 6. Notice that $f_{\mathbf{N}}(n-kq)$ is nonzero in the interval $[kq-(q/2), kq+(q/2))$.

We observe that when $q_A \rightarrow 0$, i.e., requantization is almost lossless, $D_A(q_A) \rightarrow \sigma_N^2$, the variance of N. On the other

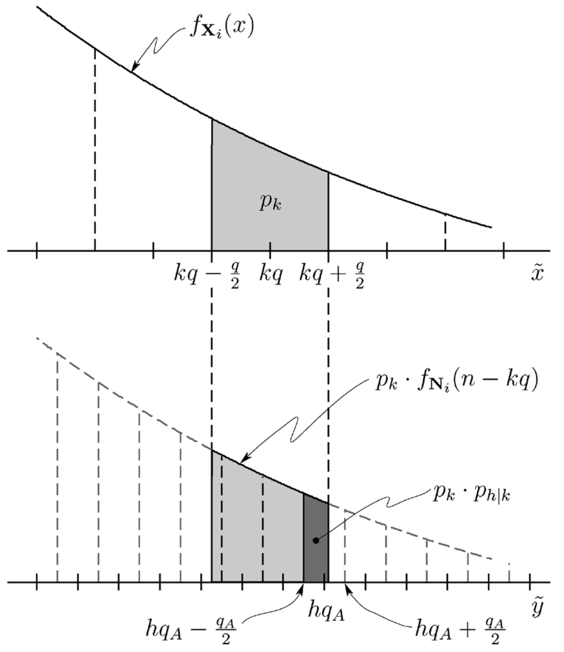


Fig. 6. Requantization of a dithered DCT coefficient (originally quantized with a quantization step size q), with a quantization step q_A. The anti-forensic dither recovers the original distribution of the coefficient $f_{\mathbf{X}}(x)$. When requantized, the noise in the k-th quantization bin is redistributed into several bins.

hand, when $q_A \to \infty$, Y is always quantized to zero. Therefore, $D_A(q_A) \to \sigma^2_{\tilde{X}}$, i.e., the variance of $\tilde{X}$. The dithering noise is canceled, i.e., $D_A(q_A) = 0$, when $q_A = (k/h)q$ *for all the values* h, k for which $p_{h|k} > 0$. This is achieved when $q_A = q$, such that $p_{h|k} = 1$, when $k = h$, and $p_{h|k} = 0$ otherwise. It can be easily seen from Fig. 6 that this corresponds to the case when all the noise N added to the coefficients in the k-th bin is reabsorbed by the quantized values $kq_A = kq$, resulting in $D_A(q_A) = 0$.

When $q_A \neq q$, requantization does not suppress distortion completely, i.e., $D_A(q_A) > 0$. Specifically, when $q_A > q$, the dithering signal is mostly canceled. Conversely, when $q_A < q$, new nonempty bins in the histogram of $\tilde{Y}$ are created, due to the dithering signal N leaking to neighboring bins (see Fig. 6). In other words, the noise N is more accurately reproduced in $\tilde{Y}$. Due to the additive nature of MSE distortion, it is possible to generalize the analysis above from individual DCT subbands to the whole image.

As an illustrative example, we measured the MSE distortion between the original JPEG-compressed image and the output of the second JPEG compression for *Lenna*. We considered the widely used JPEG quantization matrices suggested by the Independent JPEG Group [28]. Hence, quantization is adjusted by means of a 100-points quality factor $Q = 1, \ldots, 100$ that scales a template matrix to obtain the quantization steps $q_{A,i}$, $i = 1, \ldots, 64$ for each DCT subband. Similarly, recompression is driven by a quality factor Q_A. Fig. 7 illustrates the mean-square-error distortion between the recompressed image (at quality factor Q_A) and the JPEG-compressed one, when the latter was originally compressed at $Q = 35$, 60, 85. We notice a trend similar to the one predicted for each DCT coefficient subband, where the distortion is minimized when $Q_A = Q$ (thus $q_{A,i} = q_i$ for all i). The distortion is not exactly zero due to rounding and truncation of pixel values.

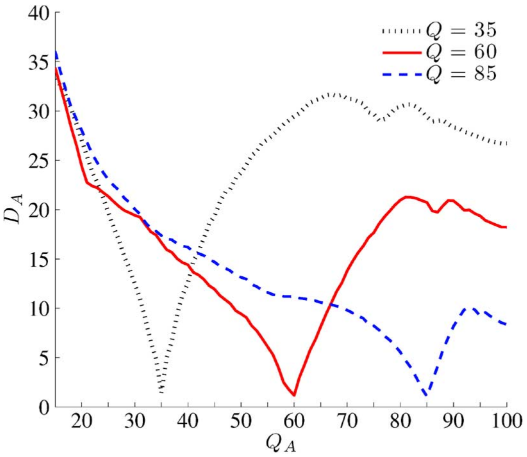


Fig. 7. MSE distortion between the *Lenna* image, JPEG-compressed at quality Q, and its recompressed version at quality Q_A. The distortion is almost equal to zero when $Q_A = Q$.

Visual inspection of the recompressed images reveals that, for $Q_A < Q$ the recompressed image does not contain traces of the dithering signal, which is mostly suppressed together with the high frequency components of the underlying image. On the other hand, for $Q_A > Q$, the dithering signal is somewhat preserved, and transformed back to the spatial domain, thus resulting in an image affected by grainy noise. This observation triggers the intuition for the detection method illustrated in the next section.

IV. Detection of JPEG Compression in the Presence of Anti-Forensics

The analysis above suggests that it is possible to identify anti-forensically dithered images by checking whether the noise introduced is annihilated after requantization. Unfortunately, in practice we do not have access to the original JPEG-compressed image in order to compute the MSE distortion after requantization. Nevertheless, we observe that the presence of the dithering signal in the spatial domain can be detected using a blind noisiness metrics. revision To this end, any metrics that can robustly measure the amount of noise present in an image could be employed. In the following, we adopt the total variation (TV) metrics [33], which is defined as the Dll_1 norm of the spatial first-order derivatives of an image. The total variation is more sensitive to small and frequent variations of pixel values due to noise, than to abrupt changes corresponding to edges. Hence, it is widely adopted as part of the objective function of optimization algorithms used for image denoising. Of course, other metrics could also be successfully employed. In Section V we will show that, for example, the mean value of the SPAM feature vector [25], can also be used. Indeed, its value is directly proportional to the amount of noise in the image, as it measures the strength of interpixel correlation.

In the following, we consider two cases. In the first one, discussed in Section IV-B, we assume that some prior knowledge

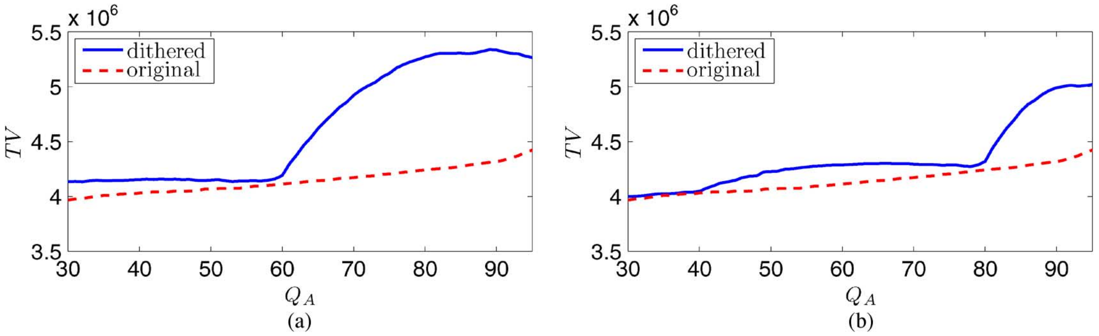


Fig. 8. Total variation (TV) as a function of the recompression quality factor Q_A, for two versions of the *Lenna* image. (a) $Q = 60$; (b) $Q = 80$.

about the original JPEG coding is available, e.g., that the original quantization matrix belongs to a family of quantization matrices corresponding to a certain JPEG implementation (e.g., the IJG implementation). In this setting, the forensic analyst can recompress the questioned image using the same quantization matrix template as the original.

In the second case, discussed in Section IV-B, we consider the more general setting in which the quantization matrix template is not available. In this case, we only make very loose assumptions about the symmetry properties that characterize typical JPEG quantization matrices.

The experimental evaluation of the two detectors on a large image dataset, including the selection of the relevant detector parameters, is postponed to Section V.

A. Known Quantization Matrix Template

In many JPEG implementations—including the IJG libjpeg software [28] and commercial photo-editing programs such as Adobe Photoshop—it is customary to use predetermined JPEG quantization matrices. The specific quantization matrix is implicitly identified when the user selects the target quality factor Q. For instance, in the IJG scheme, quantization matrices are obtained by properly scaling a template matrix, while in Adobe Photoshop, each JPEG quality factor corresponds to a specific quantization matrix stored in a lookup table.

If the forensic analyst is aware of the specific JPEG implementation that was originally used to encode the image, he can readily generate 8×8 quantization matrices $\mathbf{Q}_A$ given a (scalar) quality factor Q_A. That is, $\mathbf{Q}_A \equiv \mathbf{Q}_A(Q_A)$, where the subscript $_A$ refers to the quality factor used by the analyst. Then, he can recompress the doubted image using different analysis quality factors Q_A. For each recompressed image, the total variation $\mathrm{TV}(Q_A) \equiv \mathrm{TV}(\mathbf{Q}_A(Q_A))$ is computed. Fig. 8 shows the TV as a function of Q_A for two versions of the *Lenna* image, when the IJG scheme is employed. The dashed line corresponds to the genuine, uncompressed image. Not surprisingly, the TV increases smoothly when Q_A increases. Instead, to generate the solid line, the *Lenna* images has been compressed (at quality factor $Q = 60$) and subsequently manipulated to add a dithering signal to restore the original distribution of the DCT coefficients. The apparent slope change at $Q_A = 60$ is due to the fact that noise starts being visible when $Q_A > Q$. We observed that this behavior is general, and applies also to different kinds of visual content. Therefore, we propose to analyze the $\mathrm{TV}(Q_A)$ curve in order to devise a detector that identifies when the traces of JPEG compression have been concealed by an adversary and, in this case, to find the original quality factor Q.

In order to decide whether an image has been attacked, we consider the first order backward finite difference signal $\Delta\mathrm{TV}(Q_A)$, obtained from the total variation curve as:

$$\Delta\mathrm{TV}(Q_A) = \mathrm{TV}(Q_A) - \mathrm{TV}(Q_A - 1) \qquad (16)$$

We deem an image to have been anti-forensically attacked if

$$\left[\max_{Q_A} \Delta\mathrm{TV}(Q_A)\right] > \tau, \qquad (17)$$

where the threshold τ is a parameter that can be adjusted by the detector. In this case, we also estimate the quality factor $\hat{Q}$ of the JPEG-compressed image as:

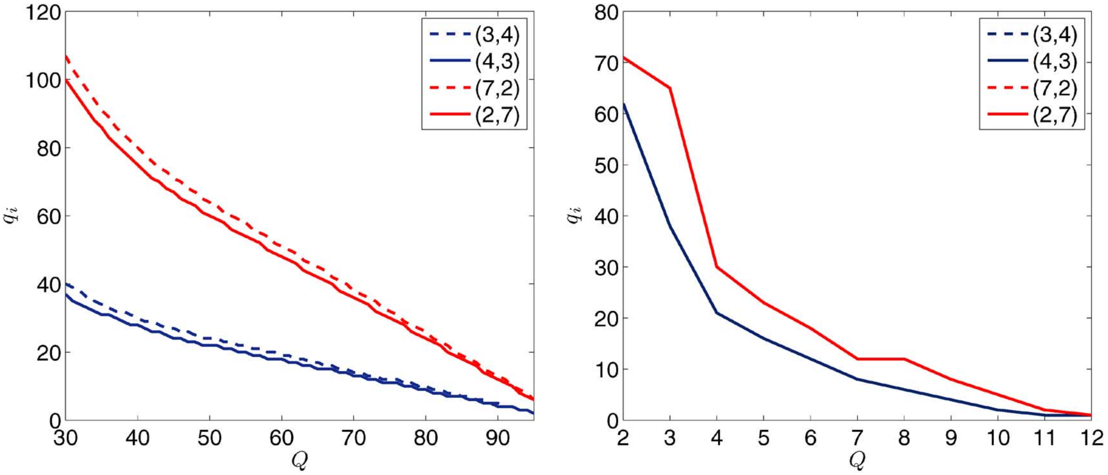
$$\hat{Q} = \left(\arg\max_{Q_A} \Delta\mathrm{TV}(Q_A)\right) - 1. \qquad (18)$$

The -1 term in (18) is to compensate for the bias introduced by the approximation of the first order derivative in (16).

B. Unknown Quantization Matrix

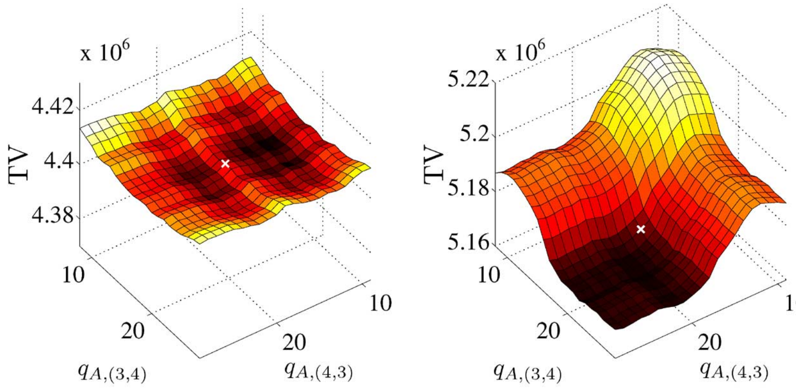
In the general case, the forensic analyst might not be aware of the JPEG implementation used to originally encode the doubted image. Unlike the case discussed in the previous section, $\mathrm{TV}(\mathbf{Q}_A)$ cannot be conveniently expressed as a function of only one scalar variable.

Indeed, the most straightforward approach would be to assume a quantization matrix template that contains the same entries for each DCT coefficient. A function $\mathrm{TV}(Q_A)$ can be obtained by scaling the values of such constant matrix, based on a scalar quality factor Q_A. However, our experiments showed that this method was ineffective in detecting traces of JPEG compression, as the function $\mathrm{TV}(Q_A)$ did not present a distinctive shape as the one illustrated in Fig. 8. This can be justified by looking at Fig. 9, which shows the dependency between the quality factor and the quantization step sizes for two pairs of DCT coefficients and two JPEG implementations. The analysis of Fig. 9 reveals the following facts: i) the dependency is almost linear, especially at higher values of the quality factor; ii) the slope depends on the DCT coefficient subbands. Therefore, scaling a constant matrix ignores the differences among



(a) IJG (b) Adobe Photoshop

Fig. 9. Quantization step size q_i as a function of quality factor Q for two pairs of DCT coefficients. (a) IJG implementation; (b) Adobe Photoshop.



(a) (b)

Fig. 10. Total variation surface for the *Lenna* image, obtained by changing the quantization steps for the DCT coefficients (3,4) and (4,3). The white cross corresponds to the true value of $(q_{A,(3,4)}, q_{A(4,3)})$. (a) Original; (b) dithered.

DCT coefficient subbands. The original quantization step sizes used to JPEG compress the doubted image correspond to different values of Q_A. Hence, the slope discontinuity in Fig. 8 would not be clearly localized at a single value of Q_A.

The quantity $\mathrm{TV}(\mathbf{Q}_A)$ can be expressed as a function of 64 variables $q_{A,i}$, $i = 1, \ldots, 64$. However, analyzing the characteristics of $\mathrm{TV}(\mathbf{Q}_A)$ in a 64-dimensional space is clearly unfeasible. Interestingly, it is possible to restrict the analysis to a two-dimensional space. Specifically, we consider a recompression scheme where the analysis quantization matrix, $\mathbf{Q}_A$, is designed in such a way that quantization affects only two DCT subbands. That is, given a pair of DCT subbands (i_1, j_1) and (i_2, j_2), we vary $q_{A,(i_1,j_1)}$ and $q_{A,(i_2,j_2)}$, whereas we set $q_{A,(i,j)} = 1$ for any $(i,j) \notin \{(i_1,j_1),(i_2,j_2)\}$.

Fig. 10 illustrates an example of $\mathrm{TV}(\mathbf{Q}_A) = \mathrm{TV}(1, \ldots, q_{A,(3,4)}, q_{A,(4,3)}, \ldots, 1)$ obtained by varying $q_{A,(3,4)}$ and $q_{A,(4,3)}$. Fig. 10(a) refers to the uncompressed *Lenna* image. Conversely, Fig. 10(b) refers to same image, but compressed with the IJG libjpeg software at quality factor $Q = 60$, and manipulated by an adversary by adding the anti-forensic dither. We observe that Fig. 10(b) exhibits a distinctive behavior with respect to Fig. 10(a).

In order to further reduce the dimensionality of the problem, we notice that JPEG quantization matrices are designed in such a way that they are approximately symmetric, i.e., $q_{(i,j)} \simeq q_{(j,i)}$. This is illustrated in Fig. 9 for two pairs of DCT coefficients. We observed this property in most commonly used JPEG quantization matrices, including those employed in IJG libjpeg, in popular photo-editing software such as Adobe Photoshop, and in digital cameras of several brands. As a result, we can further restrict search space by considering symmetric DCT subbands pairs, i.e., $(i_2, j_2) = (j_1, i_1)$, and by varying both quantization step sizes simultaneously, i.e., $q_{A,(i_1,j_1)} = q_{A,(i_2,j_2)}$. Intuitively, this corresponds to evaluating the TV function in Fig. 10 only along the diagonal.

Based on the these observations, the proposed anti-forensic detector works as follows. Each image is recompressed $(q_{max} - q_{min} + 1)$ times, by setting, at each round, $q_{A,(i_1,j_1)} = q_{A,(i_2,j_2)} = q_A$, with $q_A = q_{min}, \ldots, q_{max}$. Hence, a $(q_{max} - q_{min} + 1)$-dimensional vector $\mathrm{TV}_{(i,j)}$ is populated as

$$\mathrm{TV}_{(i,j)} = \begin{bmatrix} \mathrm{TV}(1, \ldots, q_{max}, q_{max}, \ldots, 1) \\ \mathrm{TV}(1, \ldots, q_{max} - 1, q_{max} - 1, \ldots, 1) \\ \cdots \\ \mathrm{TV}(1, \ldots, q_{min}, q_{min}, \ldots, 1) \end{bmatrix} \quad (19)$$

Note that we sort the elements of the vector $\mathrm{TV}_{(i,j)}$ in decreasing order of q_A, in such a way that the first (last) element

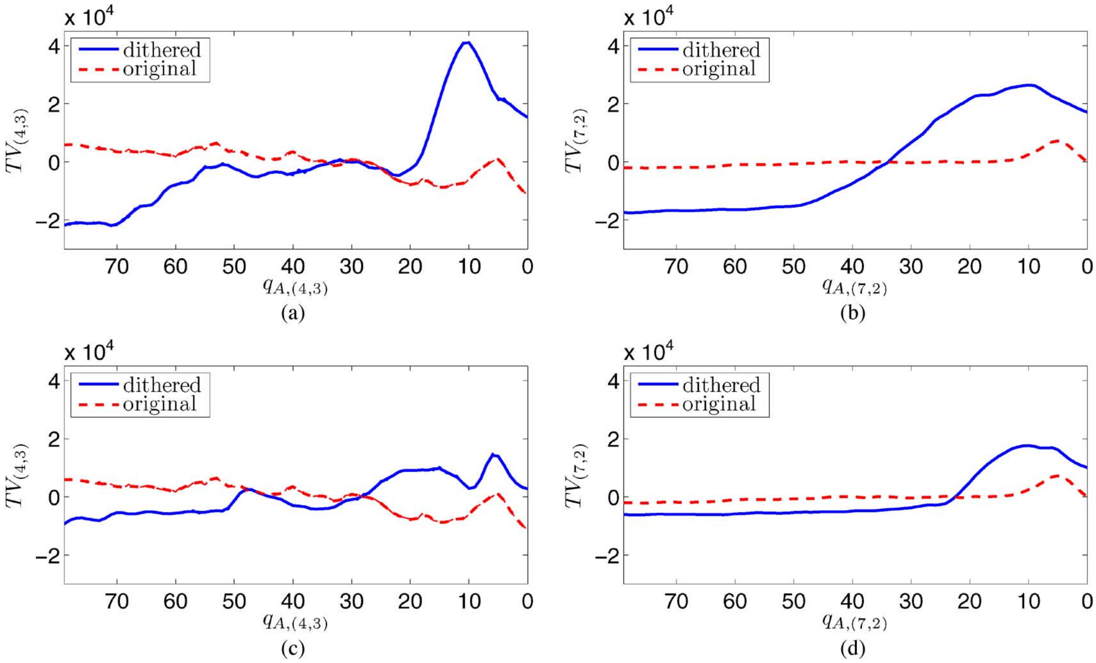


Fig. 11. Total variation $\mathrm{TV}_{(i,j)}$ as a function of q_A for two JPEG quality factors $Q = \{60, 80\}$ and two pairs of DCT coefficients $(i, j) = \{(4, 3), (7, 2)\}$. (a) $Q = 60$; (b) $Q = 60$; (c) $Q = 80$; (d) $Q = 80$.

correspond to coarser (finer) requantization, to retain the same convention adopted for the case of known matrix template.

Fig. 11 shows $\mathrm{TV}_{(i,j)}$ for two quality factors $Q = \{60, 80\}$ (using IJG implementation) and two pairs of DCT coefficients $(i, j) = \{(4, 3), (7, 2)\}$. Vectors are normalized to have zero mean for display purposes. We observe that the coefficient (4,3) leads to a vector which is noisier than the one obtained using (7,2), especially at a higher JPEG quality factor. Thus, the latter leads to a more robust detector when it comes to discriminate dithered and original images. This confirms the analysis in Section III, where we showed in Fig. 4 that the MSE introduced by dithering is significantly higher in DCT coefficient (7,2) than in (4,3).

In order to distinguish between uncompressed and dithered images, first, we compute a smoothed vector $\overline{\mathrm{TV}}_{(i,j)}$ by means of local regression using weighted linear least squares and a 2nd degree polynomial model. This operation removes noisy variations of the $\mathrm{TV}_{(i,j)}$ vector. Then, similarly to the case of known matrix template, we compute the first order derivative

$$\Delta\overline{\mathrm{TV}}_{(i,j)}(q_A) = \overline{\mathrm{TV}}_{(i,j)}(q_A) - \overline{\mathrm{TV}}_{(i,j)}(q_A - 1), \qquad (20)$$

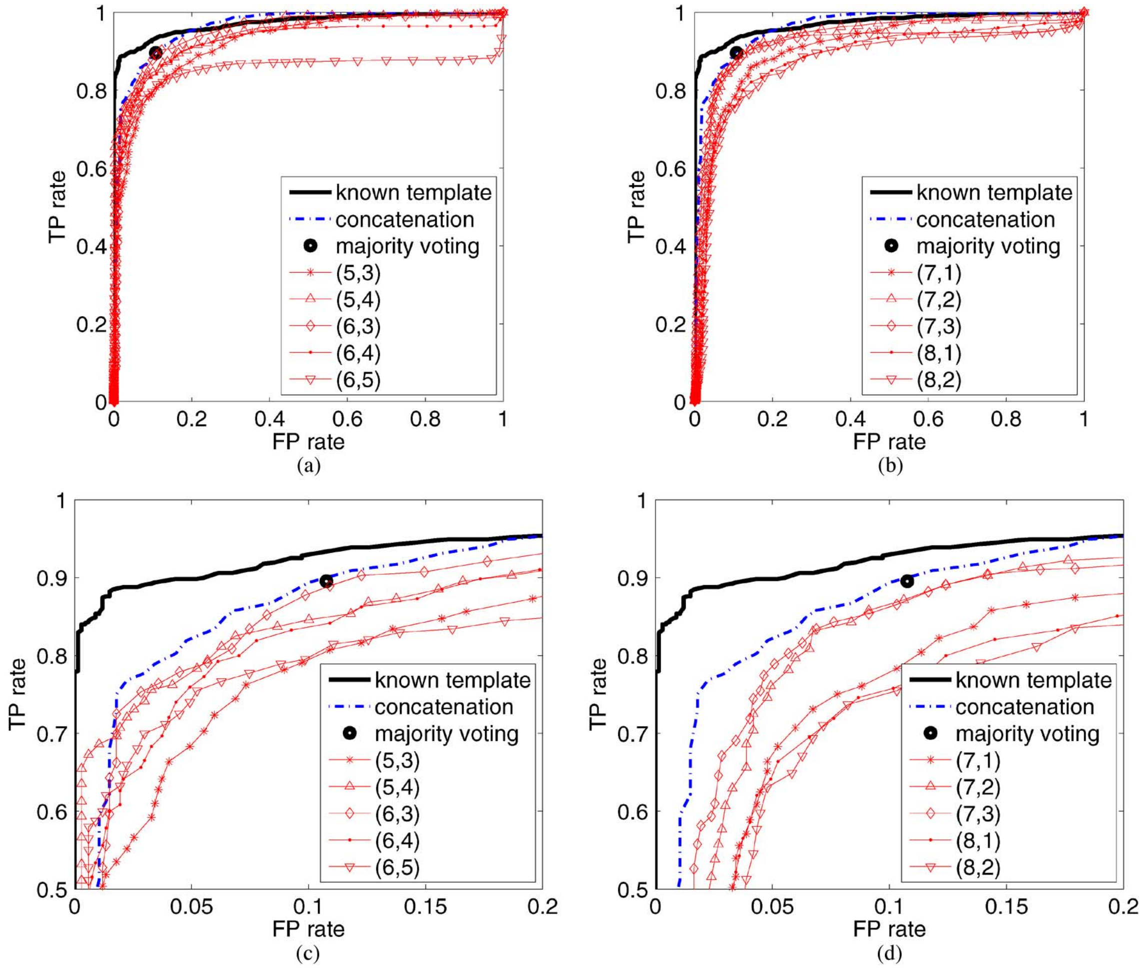
and we deem an image to have been anti-forensically attacked if

$$\left[\max_{q_A} \Delta\overline{\mathrm{TV}}_{(i,j)}(q_A)\right] > \tau_{(i,j)}. \qquad (21)$$

The detector in (20) and (21) exploits only a single pair of DCT coefficient subbands. We argue that it is possible to improve the accuracy of the detector by merging together the observations gathered from multiple DCT coefficient subbands. Therefore, we consider a set $\mathcal{C} = (i_1, j_1)\ldots., (i_C, j_C)$ of C subbands and we pursue two different information fusion approaches, which correspond to, respectively, preclassification and postclassification according to the taxonomy proposed in [34]. In the preclassification approach, we construct a vector $\Delta\overline{\mathrm{TV}}_{\mathcal{C}} = [\Delta\overline{\mathrm{TV}}^T_{(i_1,j_1)}, \ldots, \Delta\overline{\mathrm{TV}}^T_{(i_C,j_C)}]^T$ of size $(q_{max} - q_{min}) \cdot C$ obtained concatenating the vectors corresponding to the individual subbands. The detector is identical to (21), where $\Delta\overline{\mathrm{TV}}_{(i,j)}$ is replaced by $\Delta\overline{\mathrm{TV}}_{\mathcal{C}}$. In the postclassification approach, for each subband we compute a binary decision according to (21), and the final decision is obtained by majority voting among the set of C binary decisions. The accuracy of the different versions of the detector is studied in Section V.

V. Experiments

We carried out a large-scale test of the algorithms described in Section IV-A and Section IV-B on 1338 images of the Uncompressed Color Image Database (UCID) [32]. All the pictures in this dataset have a resolution of 512 × 384. Without loss of generality, we considered the luma component only. Some of the methods described in the literature require training, for which we adopted a separate dataset. The NRCS dataset [35] was obtained downloading 978 uncompressed images (raw scans of film with resolution 2100 × 1500). In order to provide a fair comparison, we resampled the images at the same resolution as the UCID dataset used for testing. We split the dataset in two sets of equal size. The first half contained images that were JPEG-compressed at a random quality factor Q using the IJG implementation. More specifically, the quality factor is uniformly sampled in the set $\{30, 40, 50, 60, 70, 80, 90, 95\}$ with probability 1/8. In order to restore the original statistics of the DCT coefficients, we added an anti-forensic dithering signal according to the method in [18]. The remaining half contained uncompressed original images.



(a)

(b)

(c)

(d)

Fig. 12. ROC curves of the proposed detector. All figures report the curves corresponding to: i) known matrix template; ii) unknown matrix template, concatenating $C = 10$ DCT subbands; iii) unknown matrix template, fusing the outputs of the individual detectors by majority voting. In addition, we show one curve for each DCT coefficient subband used by the detector. To avoid cluttering the figure, a set of five coefficients is shown in (a)–(c) and the other set in (b)–(d). (c) Zoom of (a); (d) zoom of (b).

A. Known Quantization Matrix Template

In the case of known quantization matrix template, we let the threshold τ vary to trace the receiver operating characteristic (ROC) curve shown in Fig. 12. There, the true positive (TP) rate is the fraction of JPEG-compressed images that were correctly reported to be compressed and the false positive (FP) rate is the fraction of uncompressed images that were reported to be compressed. Typically, the forensic analyst is interested to work at a low target FP rate. Fig. 12(c) illustrates a zoomed version of the ROC for FP rates in the interval [0, 0.2]. We observe that the detector reaches a TP rate above 0.89 at a FP rate as low as 0.02. Fig. 12 is obtained by considering all the images in the dataset. Thus, it does not reveal the performance of the proposed method for different values of the (unknown) quality factor Q. In order to study this aspect, Fig. 13 illustrates individual ROC curves for each value of Q. Each curve is obtained by considering a subset of the original dataset, which is constructed by taking all the images that were JPEG compressed at quality factor Q, and an equivalent number of uncompressed images selected at random, so as to obtain a balanced subset. Fig. 13 shows the excellent performance of the proposed detector when Q is in the range [30, 90]. The only critical situation is obtained when $Q = 95$. However, in the last part of this section we will show that, building a more sophisticated detector that uses the proposed $\mathrm{TV}(Q_A)$ feature as input to an SVM classifier, very good results can be achieved also when $Q = 95$.

When the matrix template is known, the forensic analyst might be also interested in determining an estimate $\hat{Q}$ of the quality factor Q originally used to compress the image. Table I reports the performance of the estimator in terms of:

- Bias

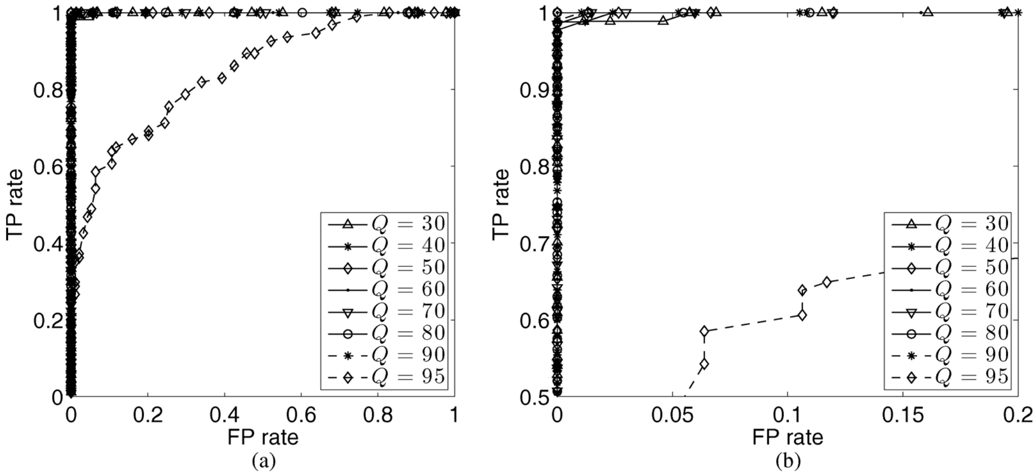
$$\varepsilon_{\mathrm{bias}} = \frac{1}{TP} \sum_{p \in \mathcal{S}} \left(\hat{Q}^{(p)} - Q^{(p)} \right) \tag{22}$$

- Standard deviation

$$\varepsilon_{\mathrm{sd}} = \sqrt{\frac{1}{TP} \sum_{p \in \mathcal{S}} \left| \hat{Q}^{(p)} - Q^{(p)} \right|^2} \tag{23}$$

- Fraction of errors above 5 units

$$\varepsilon_5 = \frac{\left| \left\{ p \in \mathcal{S}, \left| \hat{Q}^{(p)} - Q^{(p)} \right| > 5 \right\} \right|}{TP} \tag{24}$$



(a)

(b)

Fig. 13. ROC curves of the proposed detector, for the case of known matrix template. Each curve corresponds to a different value of the (unknown) quality factor Q. (b) Zoom of (a).

TABLE I
DETECTION ACCURACY. KNOWN MATRIX TEMPLATE

Q	PSNR	accuracy	$\mathcal{E}_{\text{bias}}$	$\mathcal{E}_{\text{sd}}$	$\mathcal{E}_5$
30	29.0	0.99	-12.4	1.1	1.00
40	30.8	0.99	-2.4	2.2	0.01
50	30.8	1.00	-1.8	1.9	0.08
60	32.1	1.00	-1.5	1.6	0.04
70	33.7	0.99	-1.1	3.7	0.03
80	35.0	1.00	-1.0	0.4	0.00
90	39.2	0.99	-0.3	5.0	0.01
95	43.7	0.62	12.3	12.8	0.82

where $\mathcal{S}$ denotes the set of indexes of TP images classified as true positive samples. $Q^{(p)}$ and $\hat{Q}^{(p)}$ denote, respectively, the true and estimated quality factor of the p-th image in the set $\mathcal{S}$. Note that $Q^{(p)}$ is undefined for negative samples (TN+FP) and $\hat{Q}^{(p)}$ is not computed for false negative samples. Both bias and standard deviation are within a few units when $Q \in [40, 90]$, with larger values at the extremes of the tested range, i.e., $Q = 30, 95$.

B. Unknown Quantization Matrix Template

In the case of unknown quantization matrix, we selected a set of DCT coefficient subbands $\mathcal{C} = \{(5,3), (5,4), (6,3), (6,4), (6,5), (7,1), (7,2), (7,3), (8,1), (8,2)\}$, based on the analysis in Section III. Indeed, they correspond to mid-frequency subbands in which the MSE distortion due to the addition of the dithering noise is largest (cfr. Fig. 4). We let $\tau_{(i,j)}$ vary to trace the ROC curve for each subband $(i,j) \in \mathcal{C}$. To avoid cluttering the figure, a set of five coefficients is shown in Fig. 12(a) and Fig. 12(c) and the other set in Fig. 12(b) and Fig. 12(d). In this case, we observe that the performance of the detector depends on the selected DCT subband, with the best results achieved for subbands (5,4), (6,3), (7,2) and (7,3). With respect to the case of known matrix template, a TP rate above 0.89 is reached at a FP rate equal to 0.1, confirming the fact that the knowledge of the matrix template facilitates the work of the forensic analyst.

Fig. 12 also shows the results obtained with the two information fusion approaches described in Section IV-B. As for fusion based on preclassification, i.e., concatenating the vectors obtained with all subbands, the TP rate for a target FP rate is higher than in the case of a detector based on a single subband. The only exception is at very low FP rates, where a detector based on (5,4) achieves higher TP rate. We argue that more sophisticated fusion methods (e.g., weighting the contribution of each DCT subband differently) might further improve the results, and it is left as future work. In the case of fusion based on postclassification, majority voting provides a binary decision for each image. As such, instead of the ROC curve we can only report the corresponding TP rate versus FP rate point.

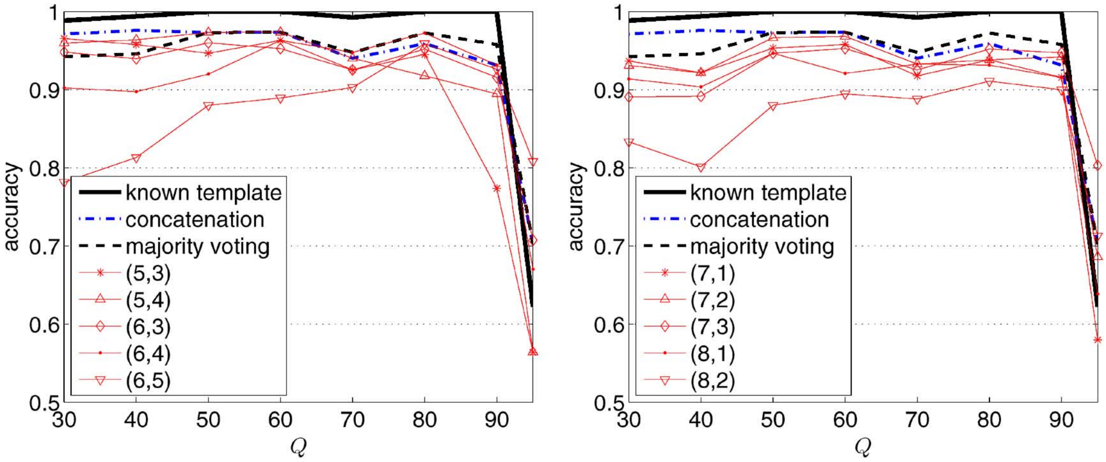
C. Detection Accuracy and Threshold Selection

The only parameter that need to be adjusted by the forensic analyst is the value of the threshold τ (or $\tau_{(i,j)}$, for the case of unknown matrix template). In order to select the optimal value of the threshold, we evaluated performance in terms of accuracy, i.e., the fraction of correct decisions taken by the detector. That is

$$\text{accuracy}(\tau) = \frac{TP(\tau) + TN(\tau)}{N}, \tag{25}$$

where TN denotes the number of true negatives, i.e., the uncompressed original images that were detected to be so, and N is the number of images in the dataset. Accuracy is a suitable metrics when the dataset is balanced, such as the one adopted in our experiments. For each value of the threshold, we computed the accuracy of the detector and we selected the threshold that maximizes accuracy over the whole dataset. Note that the optimal value of the thresholds (i.e., τ^*, $\tau^*_{(i,j)}$) has been obtained over a dataset characterized by diverse coding conditions, as the forensic analyst is typically unaware of them, so as to provide a fair evaluation.

Table II shows the overall accuracy, averaged over all JPEG coding conditions. We observe that leveraging the knowledge of the matrix template we achieve a level of accuracy equal to 0.93. When the matrix template is unknown and the detector is based on individual DCT coefficient subbands, accuracy



(a) (b)

Fig. 14. Accuracy of the proposed detector versus Q. All figures report the curves corresponding to: i) known matrix template; ii) unknown matrix template, concatenating $C = 10$ DCT subbands; iii) unknown matrix template, fusing the outputs of the individual detectors by majority voting. In addition, we show one curve for each DCT coefficient subband used by the detector. To avoid cluttering the figure, a set of five coefficients is shown in (a) and the other set in (b).

TABLE II
DETECTION ACCURACY AVERAGED OVER ALL JPEG CODING CONDITIONS (IN ORDER OF DECREASING ACCURACY)

	detector	accuracy
	known matrix template	0.93
unknown matrix template	concatenation	0.90
	majority voting	0.89
	(6,3)	0.89
	(7,2)	0.88
	(7,3)	0.88
	(5,4)	0.88
	(6,4)	0.87
	(7,1)	0.86
	(5,3)	0.85
	(6,5)	0.85
	(8,1)	0.84
	(8,2)	0.83

decreases (0.83–0.89). Both information fusion methods help bridging the gap with the case of known matrix template.

In addition, we evaluated the performance of the detectors for each value of Q. In this case, let N_Q denote the number of images in the dataset that were originally compressed with JPEG with a quality factor equal to Q. To evaluate accuracy, we set $N = 2 \cdot N_Q$ and randomly selected a subset of N_Q uncompressed original images as negative samples, in order to ensure a balanced population between positive (JPEG-compressed images) and negative (uncompressed original images) samples. Fig. 14 shows the results obtained when the quantization matrix template is either known or unknown. In the latter case, we report results obtained with detectors leveraging individual DCT subbands and by fusing the information of different subbands. The detection accuracy is only mildly dependent on the JPEG coding conditions in the [30,90] range, and it drops when $Q = 95$. Indeed, at higher quality, JPEG compression is nearly lossless (PSNR is above 46 dB), and the traces left by quantization can be easily concealed without introducing significant distortion. Excluding $Q = 95$, accuracy is above 99% for the case of known matrix template and above 95% for the case of unknown matrix template when information fusion based on majority voting is used.

D. Comparison With Steganalysis

We compared the proposed detectors with other methods described in the literature, namely detectors based on the Subtractive Pixel Adjacency Matrix (SPAM) features in [25] and the calibration features in [23]. The SPAM features were proposed to detect steganographic methods that embed in the spatial domain by adding a low-amplitude independent stego signal. Although originally designed for a different purpose, these features can be used to build a detector by considering the uncompressed original image as the cover image, and the decompressed JPEG image with added the anti-forensic dithering signal as the stego image. We extracted first order SPAM features using the default parameters suggested in [25], which leads to a 162-dimensional vector for each image. This feature vector describes the amount of interpixel correlation along different directions, and we refer the reader to [25] for details on how it is computed. In order to design a detector, a support vector machine (SVM) classifier was trained on the NRCS dataset. In order to tune the parameters of the SVM classifier (C, γ) we followed the procedure described in [25], which suggests five-fold cross-validation with an exhaustive search over a multiplicative grid sampling the parameter space. We did not consider second-order SPAM features as the feature vector has a dimension comparable to the size of the training set, thus being at risk for overfitting.

The calibration features were also originally proposed in the field of steganalysis [24] and recently adapted to detect JPEG compression anti-forensics [23]. Specifically, a single calibration feature is used as in [23], which measures the ratio of the variance of high frequency subbands. The forensic analyst crops the doubted image $\mathbf{Y}$ in the spatial domain by 4 pixels in both horizontal and vertical direction to obtain a new image $\mathbf{Z}$. Then the sample variance of 28 high frequency subbands in a set $\mathcal{C}$ is

TABLE III
ACCURACY OF THE DETECTORS VERSUS Q

	known template	detector	30	40	50	60	70	80	90	95	avg.	avg 30-90
$\mathrm{TV}(Q_A)$	y	thresh.	0.99	0.99	1.00	1.00	0.99	1.00	1.00	0.62	0.93	1.00
$\mathrm{TV}(q_{A,i})$ - concat.	n	thresh.	0.97	0.98	0.97	0.97	0.94	0.96	0.93	0.70	0.90	0.96
$\mathrm{TV}(q_{A,i})$ - voting	n	thresh.	0.97	0.98	0.97	0.97	0.94	0.96	0.93	0.70	0.90	0.96
Calibration feat.	n	thresh.	0.98	0.99	0.99	0.98	0.98	0.99	0.96	0.81	0.97	0.98
SPAM feat.	n	SVM	0.98	0.96	0.98	0.97	0.96	0.98	0.96	0.94	0.95	0.97
$\mathrm{TV}(Q_A)$	y	SVM	0.98	0.98	0.97	0.98	0.94	0.97	0.97	0.93	0.96	0.97
$\mathrm{SPAM}(Q_A)$	y	SVM	0.97	0.93	0.99	0.96	0.98	0.99	0.93	0.88	0.95	0.96
$\mathrm{PSNR}(Q_A)$	y	SVM	0.95	0.94	0.93	0.90	0.89	0.77	0.94	0.83	0.85	0.90

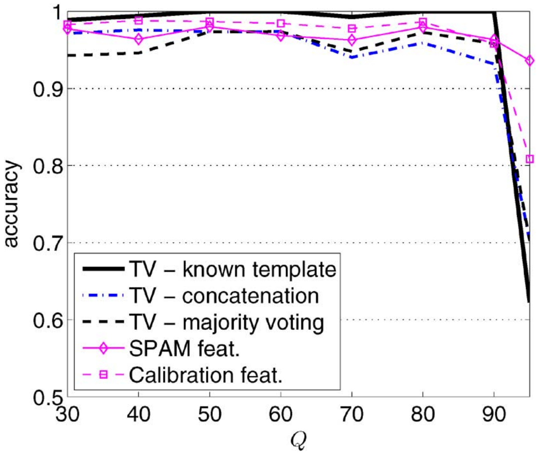


Fig. 15. Accuracy of the detectors versus Q. Comparison between the proposed methods and detectors based on: (i) SPAM features; (ii) Calibration features.

computed for both images. The calibration feature F is calculated as follows:

$$F = \frac{1}{|\mathcal{C}|}\sum_{i\in\mathcal{C}}\left(\frac{\sigma^2_{\mathbf{Z},i} - \sigma^2_{\mathbf{Y},i}}{\sigma^2_{\mathbf{Y},i}}\right) \tag{26}$$

The detector compares the value of F to a threshold τ_F. If $F \leq \tau_F$, the image is considered to be an uncompressed original.

The accuracy of the detectors for different values of the JPEG quality factor Q is illustrated in Fig. 15. In addition, Table III reports the average accuracy both over all tested values of Q and in the [30,90] interval. For each detector, we also specify if it relies on prior knowledge of the quantization template, and if it requires training. The proposed detector achieves the highest accuracy over the range [30,90], reaching almost 100%, when the quantization matrix template is known. The detector based on calibration features also gives very good results, slightly outperforming the proposed fusion methods as well as the detector based on SPAM features. The latter detector outperforms all the others when $Q = 95$, thus yielding uniform accuracy across the whole range of tested values of Q. Stimulated by this finding, we also tested another detector, which is based on the $\mathrm{TV}(Q_A)$ feature vector. Instead of the simple threshold-based detector described in Section IV, a SVM classifier was trained using the same procedure used for the detector based on SPAM features. In this case, results are very similar to those obtained for SPAM, also when $Q = 95$.

Note that the detector based on SPAM features is prone to give false positives when other kinds of noise are introduced into the image, since it only considers interpixel correlations, thus disregarding the underlying processing chain of JPEG compression anti-forensics. Using a terminology borrowed from statistics, we could say that SPAM features have a high *sensitivity* to JPEG anti-forensic dithering. That is, they can identify accurately the case in which an image has been compressed and anti-forensically dithered. However, they have a low *specificity*, since they could classify uncompressed images indeed as anti-forensically dithered. To show this, we added i.i.d. Gaussian noise to the uncompressed original images and fed these images as input to the detector. As expected, the detector based on SPAM features always reported that the images were JPEG compressed and subject to anti-forensic dithering, although they were never compressed before. Fig. 16 shows the accuracy of the compared methods for different values of Q. Notice that the accuracy of the detector based on SPAM features is close to 0.5, meaning that there are as many false positives as true positives. On the other hand, our method correctly classifies images degraded with Gaussian noise as uncompressed (although with lower accuracy than for the noiseless case), therefore achieving both high sensitivity and specificity. This is due to the fact that the proposed method takes advantage of the knowledge of the processing chain of JPEG compression and JPEG compression anti-forensics.

At the beginning of Section IV we argued that other metrics which can robustly measure the amount of noise present in an image could be employed. Hence, we tested a detector based on the $\mathrm{SPAM}(Q_A)$ feature vector. That is, instead of computing the value of the total variation, it extracts the average value of the SPAM feature vector on the output of JPEG compression at quality Q_A. Unlike the case of the $\mathrm{TV}(Q_A)$, it was not possible to define a simple detector based on a simple threshold value. Hence, we designed a detector based on a trained SVM classifier. This detector achieves a level of accuracy which is comparable to the one obtained using $\mathrm{TV}(Q_A)$, thus supporting the claim that other noisiness metrics can also be employed.

Finally, we considered a detector that uses $\mathrm{PSNR}(Q_A)$ as feature vector. The vector is obtained computing the PSNR value between the input and the output of the JPEG compression step at quality Q_A. The detector is designed by means of a SVM classifier as before. This is inspired to similar methods presented in the literature, that exploit the idempotency property of quantization to, e.g., estimate the quality factor in JPEG

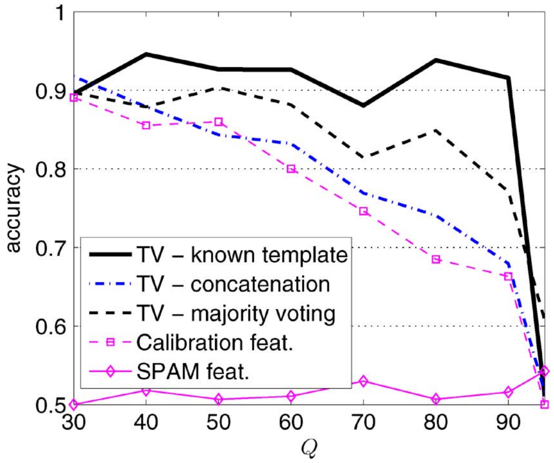


Fig. 16. Accuracy of the detectors versus Q, in the case where additive white Gaussian noise has been added to the uncompressed images. Comparison between the proposed methods and detectors based on: (i) SPAM features; (ii) Calibration features.

compressed images [20], exposing forgeries [21] and to identify the video codec [22]. Table III shows that it is generally outperformed by the other tested methods.

Overall, it is possible to observe that methods originally developed for steganalysis, i.e., SPAM and calibration features, can be effectively adopted to reveal the traces of JPEG compression anti-forensics, achieving very good results, comparable to our method, for a wide range of quality factors. However, since the proposed method is specifically tailored to detect JPEG compression in the presence of anti-forensics, it yields the following additional advantages: i) unlike the detector based on SPAM features, the proposed method does not lead to false positives when the image is degraded by noise other than anti-forensic dithering; ii) unlike the detectors based on either SPAM or calibration features, the proposed method is able to estimate the underlying JPEG quality factor (for a known template quantization matrix) or some elements of the quantization matrix, when JPEG compression is detected.

VI. Conclusions

JPEG compression leaves characteristic footprints which can be potentially exploited by the forensic analyst to perform tampering detection, source identification, etc. Recently, it has been shown that an adversary might conceal such footprints by adding a properly designed dithering noise signal in the DCT domain. The paper investigates the problem of JPEG-compression anti-forensics by showing how the forensic analyst can effectively counter the anti-forensic method originally proposed in [18]. Our analysis proves that removing traces of JPEG compression is more difficult than previously thought. Furthermore, our approach differs from conventional steganographic techniques in that it specifically targets JPEG anti-forensic dither, thus it is less prone to produce false positives when the image has been corrupted by other nonmalicious kinds of noise. In addition, if the quantization matrix is a scaled version of a known template, it is able to estimate the underlying JPEG quality factor. Future research will investigate the problem of compression anti-forensics in the field of video coding. There, motion-compensation provides a further element both the forensic analyst and the adversary can play with. In order to enable reproducible research, a software-based implementation of the forensics and anti-forensics tools described in this paper are made publicly available at www.rewindproject.eu and rewind.como.polimi.it.

References

[1] G. Valenzise, M. Tagliasacchi, and S. Tubaro, "The cost of JPEG compression anti-forensics," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, May 2011.

[2] G. Valenzise, V. Nobile, M. Tagliasacchi, and S. Tubaro, "Countering JPEG anti-forensics," in *Proc. Int. Conf. on Image Process.*, Bruxelles, Belgium, Sep. 2011.

[3] H. Farid, "Image forgery detection," *IEEE Signal Process. Mag.*, vol. 26, no. 2, pp. 16–25, Mar. 2009.

[4] S. Milani, M. Fontani, P. Bestagini, M. Barni, A. Piva, M. Tagliasacchi, and S. Tubaro, "An overview on video forensics," *APSIPA Trans. Signal Inf. Process.*, vol. 1, pp. 1–18, 2012.

[5] M. Chen, J. Fridrich, M. Goljan, and J. Lukas, "Determining image origin and integrity using sensor noise," *IEEE Trans. Inf. Forensics Security*, vol. 3, no. 1, pp. 74–90, Mar. 2008.

[6] Z. Fan and R. L. de Queiroz, "Identification of bitmap compression history: JPEG detection and quantizer estimation," *IEEE Trans. Image Process.*, vol. 12, no. 2, pp. 230–235, Feb. 2003.

[7] W. S. Lin, S. K. Tjoa, H. V. Zhao, and K. J. Liu, "Digital image source coder forensics via intrinsic fingerprints," *IEEE Trans. Inf. Forensics Security*, vol. 4, no. 3, pp. 460–475, Sep. 2009.

[8] A. C. Popescu and H. Farid, "Exposing digital forgeries by detecting traces of resampling," *IEEE Trans. Signal Process.*, vol. 53, no. 2, pp. 758–767, Feb. 2005.

[9] W. Luo, Z. Qu, J. Huang, and G. Qiu, "A novel method for detecting cropped and recompressed image block," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, Apr. 2007, vol. 2, pp. 217–220.

[10] M. C. Stamm and K. J. R. Liu, "Forensic detection of image manipulation using statistical intrinsic fingerprints," *IEEE Trans. Inf. Forensics Security*, vol. 5, no. 3, pp. 492–506, Sep. 2010.

[11] H. Farid, Digital Image Ballistics From JPEG Quantization, Dept. Comput. Sci., Dartmouth College, Tech. Rep. TR2006-583, 2006.

[12] C. Chen, Y. Q. Shi, and W. Su, "A machine learning based scheme for double JPEG compression detection," in *Proc. Int. Conf. Pattern Recognition*, 2008, pp. 1–4.

[13] J. Lukáš and J. Fridrich, "Estimation of primary quantization matrix in double compressed JPEG images," in *Proc. Digital Forensic Research Workshop*, 2003.

[14] F. Huang, J. Huang, and Y. Q. Shi, "Detecting double JPEG compression with the same quantization matrix," *IEEE Trans. Inf. Forensics Security*, vol. 5, no. 4, pp. 848–856, Dec. 2010.

[15] J. Fridrich, D. Soukal, and J. Lukás, "Detection of copy-move forgery in digital images," in *Proc. Digital Forensic Research Workshop*, Cleveland, OH, Aug. 2003.

[16] S. Bayram, H. T. Sencar, and N. Memon, "A survey of copy-move forgery detection techniques," in *Proc. IEEE Western New York Image Processing Workshop*, Rochester, NY, Oct. 2008.

[17] A. Bianchi, T. De Rosa, and A. Piva, "Improved DCT coefficient analysis for forgery localization in JPEG images," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, Prague, Czech Republic, May 2011.

[18] M. C. Stamm, S. K. Tjoa, W. S. Lin, and K. J. R. Liu, "Anti-forensics of JPEG compression," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, Dallas, TX, Apr. 2010.

[19] M. C. Stamm, S. K. Tjoa, W. S. Lin, and K. J. R. Liu, "Undetectable image tampering through JPEG compression anti-forensics," in *Proc. Int. Conf. Image Process.*, Hong Kong, Sep. 2010, pp. 2109–2112.

[20] W. Luo, Y. Wang, and J. Huang, "Security analysis on spatial 1 steganography for JPEG decompressed images," *IEEE Signal Process. Lett.*, vol. 18, no. 1, pp. 39–42, Jan. 2011.

[21] H. Farid, "Exposing digital forgeries from JPEG ghosts," *IEEE Trans. Inf. Forensics Security*, vol. 4, no. 1, pp. 154–160, Mar. 2009.

[22] P. Bestagini, A. Allam, S. Milani, M. Tagliasacchi, and S. Tubaro, "Video codec identification," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, Mar. 25–30, 2012, pp. 2257–2260.

[23] S. Y. Lai and R. Böhme, "Countering counter-forensics: The case of JPEG compression," in *Information Hiding*. New York: Springer, 2011, pp. 285–298.

[24] J. Fridrich, "Feature-based steganalysis for JPEG images and its implications for future design of steganographic schemes," in *Proc. Inf. Hiding Workshop, Springer LNCS*, 2004, pp. 67–81.

[25] T. Pevny, P. Bas, and J. Fridrich, "Steganalysis by subtractive pixel adjacency matrix," *IEEE Trans. Inf. Forensics Security*, vol. 5, no. 2, pp. 215–224, Jun. 2010.

[26] T. Bianchi and A. Piva, "Image forgery localization via block-grained analysis of JPEG artifacts," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 3, pp. 1003–1017, Jun. 2012.

[27] S. Tubaro, S. Milani, and M. Tagliasacchi, "Discriminating multiple jpeg compression using first digit features," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, Mar. 25–30, 2012.

[28] The Independent JPEG Group [Online]. Available: http://www.ijg.org/

[29] *Recommendation T.81: Digital Compression and Coding of Continuous-Tone Still Images*, ISO/IEC 10918-1, ITU-T, 1991.

[30] E. Y. Lam and J. W. Goodman, "A mathematical analysis of the DCT coefficient distributions for images," *IEEE Trans. Image Process.*, vol. 9, no. 10, pp. 1661–1666, Oct. 2000.

[31] J. R. Price and M. Rabbani, "Biased reconstruction for JPEG decoding," *IEEE Signal Process. Lett.*, vol. 6, no. 12, pp. 297–299, Dec. 1999.

[32] G. Schaefer and M. Stich, "UCID: An uncompressed colour image database," in *Proc. SPIE: Storage and Retrieval Methods and Applications for Multimedia*, 2004, vol. 5307, pp. 472–480.

[33] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1–4, pp. 259–268, 1992.

[34] A. K. Jain, K. Nandakumar, and A. Ross, "Score normalization in multimodal biometric systems," *Pattern Recognit.*, vol. 38, no. 12, pp. 2270–2285, 2005.

[35] NRCS Photo Gallery [Online]. Available: http://photogallery.nrcs.usda.gov/

Giuseppe Valenzise (S'07–M'11) was born in 1982. He received the Master degree (2007, *cum laude*) in computer engineering and the Ph.D. degree in electrical engineering and computer science (2011), from the Politecnico di Milano, Italy.

From July 2011 to September 2012 he was a postdoc researcher at Telecom ParisTech, Paris, France. He is currently a permanent CNRS researcher with the Laboratoire Traitement et Communication de l'Information (LTCI—UMR 5141) at Telecom Paristech. His research interests include single and multiview video coding, nonnormative tools in video coding standards, multimedia forensics, video quality monitoring, and applications of compressive sensing.

Marco Tagliasacchi (S'05–M'06) is currently Assistant Professor with the Dipartimento di Elettronica e Informazione—Politecnico di Milano, Italy. He received the Laurea degree (2002, *summa cum Laude*) in computer engineering and the Ph.D. degree in electrical engineering and computer science (2006), both from Politecnico di Milano.

He was a visiting academic at the Imperial College London (2012) and visiting scholar at the University of California, Berkeley (2004). His research interests include multimedia communications (coding, quality assessment), multimedia forensics, and information retrieval. He coauthored more than 100 papers in international journals and conferences, including award winning papers at MMSP 2012, ICIP 2011, MMSP 2009, and QoMex 2009. He has been actively involved in several EU-funded research projects. He is currently cocoordinating two ICT-FP7 FET-Open projects (REWIND and GreenEyes).

Dr. Tagliasacchi is an elected member of the IEEE MMSP Technical Committee for the term 2009–2012. He serves as Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGIES (2011 best AE award) and *APSIPA Transactions on Signal and Information Processing*. He served in the Organizing Committee of the ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC) 2009 and Digital Audio Effects (DAFx 2009). He will be General co-Chair of IEEE Workshop on Multimedia Signal Processing (MMSP 2013, Pula, Italy) and Technical Program Coordinator of IEEE International Conference on Multimedia & Expo (ICME 2015, Turin, Italy).

Stefano Tubaro (A'02–M'02) was born in Novara in 1957. He completed his studies in Electronic Engineering at Politecnico di Milano, Italy, in 1982.

He then joined the Dipartimento di Elettronica e Informazione of Politecnico di Milano, first as a researcher of the National Research Council, and then (in November 1991) as an Associate Professor. Since December 2004 he has been appointed Full Professor of Telecommunication at the Dipartimento di Elettronica e Informazione of Politecnico di Milano, where he became the Coordinator of the Telecommunications Section of the Dipartimento di Elettronica e Informazione in 2008. His current research interests are on image and video analysis for the geometric and radiometric modeling of 3-D scenes; advanced algorithms for video coding and sound processing. He authored over 150 scientific publications on international journals and conferences. He also coauthored two books on digital processing of video sequences. He coordinates the research activities of the Image and Sound Processing Group (ISPG) at the Dipartimento di Elettronica e Informazione of the Politecnico di Milano, which is involved in several research programs funded by industrial partners, the Italian Government, and by the European Commission. Moreover, he is (and has been) involved in many evaluation and concertation activities for European projects. He has also been part of the IEEE Multimedia Signal Processing Technical Committee (2005–2009), and from January 2012 he is a member of the IEEE SPS Image Video and Multidimensional Signal Processing Technical Committee. He has been guest editor of several special issues published by EURASIP journals; he also has been among the organizers of a number of international conferences: IEEE MMSP 2004/2013, IEEE ICIP 2005, IEEE AVSS 2005/2009, and IEEE ICDSC 2009, to mention a few. He is currently an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING.